EXPLORING KEY VARIABLES IN WIND TURBINE POWER CURVE MODELING

A Thesis

by

DAVID MATTHEW PEREZ

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Yu Ding
Committee Members,	Li Zeng
	Mladen Kezunovic
Head of Department,	Mark Lawley

August 2018

Major Subject: Industrial Engineering

ABSTRACT

Though substantial evidence has shown the importance of wind speed and direction in modelling a wind turbine's power curve, there remains uncertainty as to whether other variables would improve modelling efforts and which specific variables those would be. This present work expands upon prior research using the additive multiplicative kernel (AMK) technique to explore the use of additional variables. The experimental methodology involves arriving at power estimates by treating a year's amount of wind turbine data as a learning problem. Different combinations of variables are investigated and compared in terms of error reduction on the testing set using root mean square error and mean average error. Discussion on the best sets of variables combinations are presented to gain insight as to why certain variables lead to greater error reduction and whether they are likely to be included in different sets of data.

Two categories of variables emerge from the research. The first includes variables that are typically recorded in field operations including time, turbulent intensity, and the standard deviation of wind direction. Time and turbulent intensity are shown to offer promising results. The next set includes variables that measure how much wind speed and direction vary as height varies. This is especially of interest as turbine size has increased substantially over recent years. In particular, the rotor equivalent wind speed neatly captures the variation of wind speed and direction across the length of a turbine's rotor in a single value. Using this parameter with AMK leads to significant prediction error reduction, making a strong case to include it in modeling the power curve. As will be discussed, doing so proves to be a better alternative than current industry practice.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Ding, and my committee members, Dr. Zeng and Dr. Kezunovic for their guidance in making this work of research possible. Dr. Ding's support enabled me to fully focus on classes and research. Support also thankfully came from the Industrial & Systems Engineering Department in my time as a teaching assistant.

I owe a tremendous debt of gratitude to Texas A&M University for their generous financial support in the form of the Graduate Diversity Fellowship. Without the fellowship, it is doubtful if I would have been able to pursue a Master's Degree, so I will be eternally grateful to the University for giving me this opportunity of a lifetime.

I would also like to thank my cat for his companionship throughout my studies. Finally, thanks to my mother and father for all they have done to raise me to be who I am today. Their constant encouragement helped me reach my goals, and I could not be more thankful to have them in my life.

CONTRIBUTORS AND FUNDING SOURCES

This work was supported by a thesis committee consisting of Professor Ding and Professor Zeng of the Department of Industrial and Systems Engineering and Professor Kezunovic of the Department of Electrical and Computer Engineering.

The data analyzed was provided through the Power Curve Working Group. All other work conducted for the thesis was completed by the student independently.

Graduate study was supported by a fellowship from Texas A&M University. Additional support was also given by a partial research assistantship from Texas A&M Energy Institute.

TABLE OF CONTENTS

Page

ABSTRACT ii

ACKNOWLEDGEMENTS iii

CONTRIBUTORS AND FUNDING SOURCES iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

LIST OF TABLES viii

1. INTRODUCTION 1

1.1 Motivation 1
1.2 The Power Curve 2
1.3 Problem Statement 5

2. LITERATURE REVIEW 6

2.1 Binning Method 6
2.2 Rotor Equivalent Wind Speed Model 6
2.3 Additive-Multiplicative Kernel-based Power Curve 8
2.3.1 Algorithm 8
2.3.2 Bandwidth Selection 10

3. RESEARCH OUTLINE 12

3.1 Datasets 12
3.1.1 Description 12
3.1.2 Quantities of Interest 13
3.2 Objective and Performance Metrics 22
3.3 Outline of Numerical Analysis Procedure 24
3.3.1 Overall Description 24
3.3.2 Code Description 25
3.3.3 Sequence of Trials 25

4. RESULTS 29

4.1 Discussion 29
4.1.1 Gradient Trials 29
4.1.2 Miscellaneous Trials 30

4.1.3 Combined Trials 31
4.1.4 Ranked Trials 32
4.2 Predicted Power Curve Examples 33

5. CONCLUSION 37

REFERENCES 38

LIST OF FIGURES

Page

Figure 1: A Nominal Power Curve 3

Figure 2: REWS Parameters 7

Figure 3: Wind power output vs. hub height wind speed.....13

Figure 4: Wind power output vs. rotor equivalent wind speed 14

Figure 5: Wind power output vs. wind direction 15

Figure 6: Wind power output vs. standard deviation of wind direction.....16

Figure 7: Wind power output vs. turbulent intensity 18

Figure 8: Wind power output vs. clock time.....19

Figure 9: Wind power output vs. exponent of wind shear 20

Figure 10: Wind power output vs. wind veer.....21

Figure 11: Estimated vs. actual power curve - (V_R, D, A, I, T) model 34

Figure 12: Estimated vs. actual power curve - (V_R, D, A, I, T, S_a, E_a) model 35

Figure 13: Estimated vs. actual power curve - (V_R, D, A, T) model.....35

Figure 14: Estimated vs. actual power curve - best models 36

LIST OF TABLES

Page

Table 1: Gradient Trials - Set 1 29

Table 2: Gradient Trials - Set 2 30

Table 3: Gradient Trials - Set 3 30

Table 4: Miscellaneous Trials - Individual Variable Effects 31

Table 5: Miscellaneous Trials - Combined Variable Effects 31

Table 6: Combined Trials - Set 1 32

Table 7: Combined Trials - Set 2 32

Table 8: Combined Trials - Set 3 32

Table 9: Ranked Models – REWS 33

Table 10: Ranked Models – HHWS 33

1. INTRODUCTION

The purpose of this thesis is to document the research process and findings as part of the fulfilment to complete a Master of Science degree (thesis-option) in Industrial and Systems Engineering. Toward that end, the problem addressed is outlined, along with the reasons why this is a relevant issue. The body of literature consulted is discussed to illustrate the salient points in the field as well as to demonstrate the novelty of the present work. A description of the research methodology then follows, including the set of deliverables as well as a description of the data used. With the methodology established, the results are then given with accompanying discussion. Finally, the findings and take-away messages of the research are summarized in the conclusion.

1.1 Motivation

Wind energy stands to revolutionize the energy landscape by providing an alternative to fossil fuels. Wind farms already produce over 80 GW of energy in the US alone, and the Department of Energy (DOE) has put forward scenarios that the industry may triple that capacity by 2030. However, barriers still remain that limit the development of new wind projects. One longstanding concern is the uncertainty surrounding how much power a given turbine will produce. Analysts quantify the relationship between power production and ambient conditions through the power curve.

The estimated power is used primarily in performance monitoring and energy forecasting (Uluyol, Parthasarathy, Foslien, and Kim, 2011; Giebel, Brownsword, Kariniotakis, Denhard, and Draxl, 2011). Performance monitoring is an essential task in

wind fleet optimization that aims to maintain acceptable performance by addressing potential issues that can be detected by a turbine producing less power than expected. Meanwhile, energy forecasting is critical to the broader utility industry as energy traders and grid operators make decisions based on this information. By reducing the uncertainty around the power curve, wind turbine companies will unlock efficiencies both in-house and for their customers. The combination of better performance and more reliable forecasts renders the technology more secure for investment, and should prove beneficial for the growth of the industry. As such, a better model to arrive at the power curve is of great interest and is explored in this research.

1.2 The Power Curve

The power curve represents the relationship between power output of a wind turbine and speed as illustrated in Figure 1. A few features of the curve bear describing. A cut-in speed, V_{ci}, must be first reached before any power is produced. A sigmoidal relationship between power output and speed then ensues until the rated speed, V_r, is reached along with the rated power of the turbine, y_r. At this point, power production plateaus to become constant, as the turbine's control mechanism maintains the rate of rotation to prevent damage to its internal components. Power production ceases at the cut-out speed, V_{co}, to prevent damage from strong winds. Manufacturers typically market their turbines with a nominal power curve, which developers and operators can use for planning purposes. Once in the field however, a turbine rarely produces the energy predicted by this nominal power curve, though it will follow its trend.

While wind speed has long been recognized as the primary variate of interest, other environmental variables like wind direction and air density among others also impact power production. These additional factors complicate the idealized version of the power curve provided by manufacturers, which may only consider one set of conditions as speed varies.

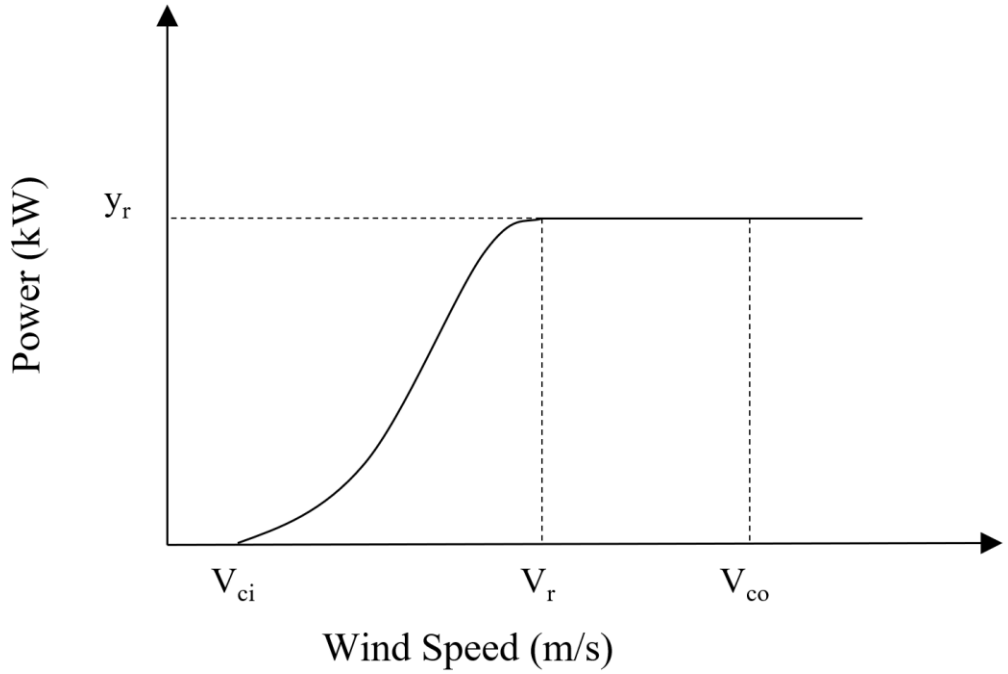

Figure 1: A Nominal Power Curve. Adapted from (Lee, Ding, Genton, and Xie, 2015).

As a result, relying on the nominal power curve given by the manufacturer alone leads to substantial error in power prediction. Further complicating the matter, no equation exists that expresses power as a closed-form analytic function of speed and these additional factors. Equation 1 gives the general relationship, but the power coefficient, C_p, cannot be expressed analytically.

$$y = \frac{1}{2} \cdot C_p(\beta, \lambda) \cdot \rho \cdot \pi R^2 \cdot V^3 \qquad (1)$$

To define the terms, y is the power production, C_p is the power coefficient, β is the blade pitch angle, λ is the turbine's tip speed ration, ρ is the air density, R is the rotor's radius, and V is the inflow speed (Belghazi and Cherkaoui, 2012). This physics-based power production equation demonstrates the primacy of speed, the influence of other factors, and the interaction effects among factors.

To address the lack of an analytical expression for wind power, industry practitioners have developed site-specific statistical models based on available historical data. These models can then be used to predict power production as conditions in the field vary. Because these models take into account the turbine's field data, the prediction results represent a reduction in prediction error compared to the idealized power curve from the manufacturer.

The most popular approach is the binning method (IEC, 2005). Given in a standard from the International Electrotechnical Commission (IEC), this method lays out a nonparametric calculation to produce a turbine's power curve, by splitting the range of speed values into bins. Subsequently averaging the power in each bin captures the power expected in the interval. Though this method offers an improvement over the manufacturer's nominal power curve, two issues are immediately evident in the approach. First, it primarily considers speed alone, though with a density correction factor. Doing so ignores the other factors understood to also influence power production. Secondly, while simple to implement, the method suffers from high error compared to competing algorithms proposed in academia. From this pair of issues, it is clear that a method that takes into account relevant factors while remaining straightforward to implement would be useful to industry analysts.

1.3 Problem Statement

Considerable effort has been expended to better understand a wind turbine's power curve. Two issues immediately stand out from the previous section however. The first is the uncertainty surrounding the choice of what variables, other than wind speed, to include in the statistical models used to arrive at an estimated power curve. Secondly, the choice of model is far from arbitrary, and a number of methods have been explored as alternatives to the binning method. It is the goal of the present work to expand on prior research on the Additive Multiplicative Kernel (AMK) model (Lee, Ding, Genton, and Xie, 2015) to provide an empirical study of the factors most pertinent to the power curve.

Beyond a few factors, there is little consensus as to which factors most strongly impact a turbine's power production. Speed, wind direction, and air density are typically considered, but other factors like wind shear, wind veer, and turbulent intensity can be relevant, too, but may not have been incorporated. Further, the extent to which each of these factors reduces error in prediction remains to be studied. Because the AMK method has previously demonstrated significant error reduction while retaining simplicity in implementation, it is the aim of this study to apply AMK to evaluate combinations of explanatory variables that lead to improved power predictions. With publicly available wind turbine datasets, the author implements AMK to find an improved selection of environmental variables.

2. LITERATURE REVIEW

2.1 Binning Method

Industry practitioners largely rely on the binning method to arrive at an estimate of the power curve. As a nonparametric method, it makes use of data to make local approximate estimates. Though this method produces satisfactory results, it exclusively makes use of wind speed to predict power, though an air density correction can be added. To briefly describe it, the binning method first splits the speed domain into a discrete number of bins. An average in each of these bins is then calculated from the available power-speed data pairs in each bin, following Equation 2. If given wind speed falls into the jth bin, its wind power, $y_{i,j}$ is used to calculate the average power in that bin, y_j.

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j} \quad for\ j = 1,2,\dots,J \qquad (2)$$

2.2 Rotor Equivalent Wind Speed Model

When a single value of speed is used for a power curve, it typically refers to the hub height wind speed (HHWS). This is the case because historically and practically, obtaining additional measurements at different heights was difficult, and wind measurements other than at the hub height were rarely available. With an increasing size in rotors in recent years however, it can no longer be assumed that the speed experienced at the hub height represents the speed throughout the area covered by the turbine's rotor. Advances in LIDAR (light detection and ranging) technology have also rendered collection of this data more readily accomplished.

The rotor equivalent wind speed (REWS) model addresses this issue by considering wind speed and direction throughout the area swept by the turbine's blades (Scheurich et al., 2016). Because the energy captured by a turbine corresponds to the area through which the blades rotate, REWS proposes slicing that circle into nearly equal areas, as illustrated below in Figure 2. REWS itself is defined in Equation 3.

$$V_R = \frac{V_H}{U(z_H)} \sqrt[3]{\frac{1}{A}\sum_{i=1}^{n} A_i[U(z_i)\cos((\theta(z_i) - \theta(z_H)]^3} \qquad (3)$$

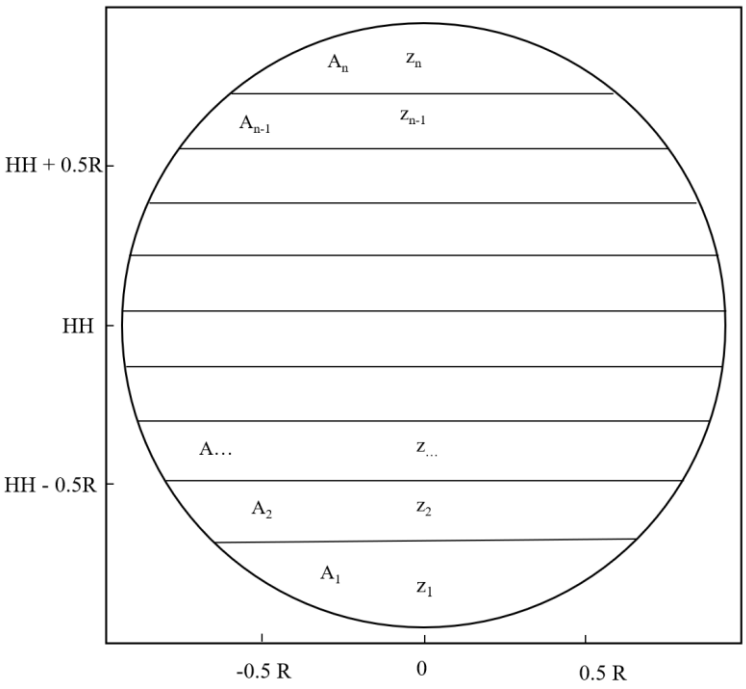

Figure 2: REWS Parameters. Adapted from (Scheurich et al., 2016).

The area, A_i, of each slice is used as a weight on the speed and wind direction passing through the height, z_i, of the slice. The speed and direction at each is aggregated into a single rotor equivalent wind speed. R represents the radius of the area swept by the blade, and z_H is the hub height. These values are used to evaluate REWS, denoted by V_R,

where $U(z)$ is the speed and $\theta(z)$ is direction at height z as measured by LIDAR. V_H in this equation refers to the hub height wind speed as measured by a cup anemometer.

The advantage of the REWS model is immediately clear in that more information is included. Less obvious is that REWS functionally takes wind shear and veer into account by considering wind speed and direction at different heights. REWS therefore has an advantage over HHWS by making use of these factors in its definition.

2.3 Additive-Multiplicative Kernel-based Power Curve

2.3.1 Algorithm

The additive multiplicative kernel model is a nonparametric, data-driven model capable both of conditional density estimation as well as point estimation. Due to the nature of the present work, it is exclusively used to arrive at point estimates for energy production dependent on explanatory variables. The ultimate goal therefore is to find the expected value of some response as expressed below, where y is power, and x is a vector of explanatory variables. Equation 4 shows this conditional expected value.

$$E[(y|x)] \qquad (4).$$

Per the method of conditional kernel density estimation (Rosenblatt 1969; Hyndman, Bashtannyk, and Grunwald, 1996) the above can be estimated as follows by m in Equation 5.

$$\hat{m}(x) = \int y\hat{f}(y|x)dy = \sum_{i=1}^{N} w_i(x)y_i \qquad (5)$$

Where

$$w_i = \frac{K_{h_x}(||x - x_i||)}{\sum_{i=1}^{N} K_{h_x}(||x - x_i||)} \qquad (6)$$

And

$$K_{h_x}(||\boldsymbol{l}||) = K_{h_x}(||l_1||)K_{h_x}(||l_2||) \dots K_{h_x}(||l_q||) \qquad (7)$$

The formulation for the mean conditional density estimator, m, does not require explicit knowledge of the conditional density function itself. Instead, the estimate is built on kernel functions of the explanatory variables acting as weights on historical data in y_i as expressed in the third term of Eqn. (5). Equations 6 and 6 show the weights and multivariate kernel function respectively. N datapoints are used in this definition, with q explanatory variables. The kernel function is typically Gaussian for most variables, but the von Mises kernel is used for wind direction because of its circular nature. As proposed by Lee et al. (2015), the mean conditional estimator can be expressed through Equation 8.

$$\hat{m} = \frac{1}{q-2}\left[\hat{m}(x_1, x_2, x_3) + \cdots + \hat{m}(x_1, x_2, x_q)\right] \qquad (8)$$

As *such*, the estimator is the average of q additive terms, each of which features an estimator using three explanatory variables at a time. The first two explanatory variables are always wind speed and wind direction, because these have been shown to be the most influential. The algorithm's principle advantage is its inherent multiplicative, nonlinear relationship between the response and multiple explanatory variables, that closely follows the formulation for power production seen Eqn. (1). It also scales with additional explanatory variables well, because each additive term only uses three variables at a time. This means that even with a relatively small dataset, the analysis undertaken will consider

at most three dimensions. Finally, a relatively quick runtimes make it a feasible choice for industry practitioners, who may not possess dedicated computing resources or the time for calculations requiring fast lead times.

2.3.2 Bandwidth Selection

As in most learning algorithms, the AMK method requires specification of parameters. The multivariate kernel functions include as many bandwidth constants as there are explanatory variables. The bandwidth constants can be explicitly represented in Equation 9. Each h in the denominator corresponds with each of the univariate kernel functions in the right-hand side of Equation 9.

$$K_{h_x}(||\boldsymbol{l}||) = \frac{1}{h_1 h_2 \dots h_q} K(||l_1||) K(||l_2||) \dots \mathrm{K}\left(\left|\left|l_q\right|\right|\right) \qquad (9)$$

The purpose of the bandwidth constants is to control the smoothness of the resulting estimates of the AMK method. Small values lead to small fluctuations from overfitting, while large values suffer the opposite problem of excessive smoothing from underfitting. Though techniques exist to find optimal bandwidth for all q explanatory variables in any given AMK model, this research instead chooses to use a heuristic algorithm from a preexisting R package in order to select the bandwidth (Ruppert, Sheather, and Wand, 1995). As an additional note, the heuristic algorithm is unable to arrive at a solution for explanatory variables that feature gaps in data when plotted against power. The figure on page 15 in the next section illustrates this issue with the plot of power against wind direction. Clear intervals of data emerge along with gaps for which there is no data. In this case, it is necessary to split the domain into intervals and proceed with the heuristic algorithm to find a separate bandwidth for each interval. Once this is

done, the appropriate bandwidth is inputted into the AMK estimate according to which interval the test set's wind direction value falls in.

3. RESEARCH OUTLINE

This section details how the research process is approached. First the dataset from the Power Curve Working Group (PCWG, 2018) is described to illustrate the input data. Each variable is described in detail, and plotted against power in the case that some noticeable trend emerges. Next, the overall objectives as well as what the deliverables include are given. The criteria to compare error is defined which will determine what model performs best. Following that is a description of the experimental setups used to generate the data itself. The guiding principles on the sequence of results are given, to explain why certain subsets of variables are included in the subsequent results section.

3.1 Datasets

3.1.1 Description

The PCWG represents a variety of stakeholders in the wind industry and has met regularly since 2012 to address many of the concerns expressed in the introduction regarding uncertainty in using power curves. Though the group has taken a different approach than that presented here, they publicly share wind data sets to encourage exploration of alternative methods to predict power production. The dataset used in this research features ~10,000 rows of data, with 29 explanatory variables and a response variable. The explanatory variables include typical features like HHWS, wind direction, density, turbulence intensity, and time. The use of LIDAR at this particular site also gives wind speed and wind direction at several different heights, which is essential in computing

REWS. A more complete description of the response and explanatory variables follows, along with preliminary data analysis relating each variate to the response.

3.1.2 Quantities of Interest

Power: P

The power generated by the wind turbine is treated as the response in this learning problem. The generated power roughly follows the power curve introduced earlier, and can be analytically defined from Eqn. (1). Because no complete analytical expression exists, practitioners instead rely on learning methods to estimate power production based on explanatory variables. The specific explanatory variables considered in this research work are given below.

Hub Height Wind Speed: V_H

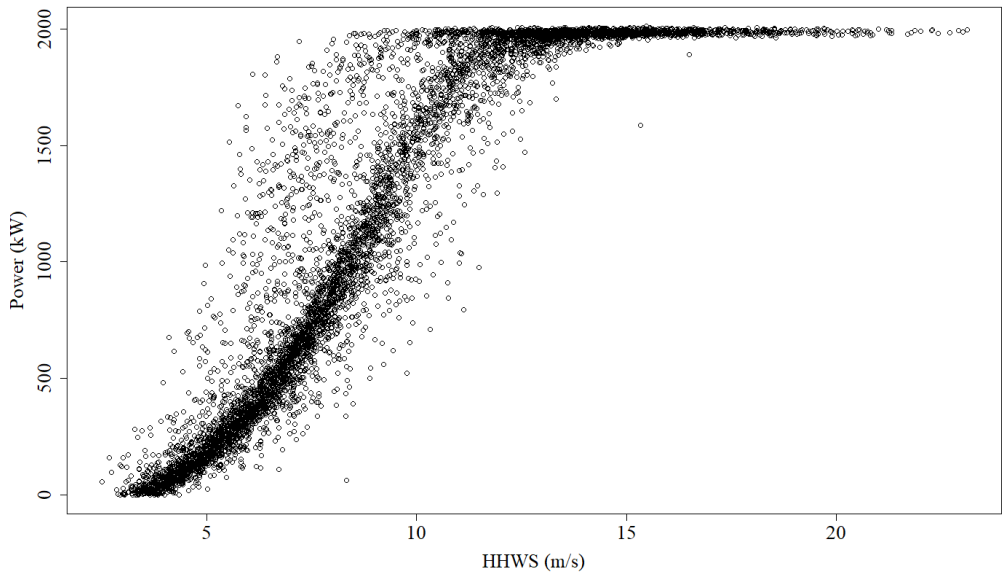

Figure 3: Wind power output vs. hub height wind speed

The speed of wind recorded at the center of the circular area swept by the wind turbine blades. Wind speed is the strongest predictor of energy production, and plotting power against wind speed outputs an approximation to the power curve as seen in Figure 3. Note that many points lie outside the curve used to describe the power curve in the ideal case. By including more of the variables discussed below, this learning problem hopes to minimize the discrepancy between the ideal case and actual behavior.

Rotor Equivalent Wind Speed: V_R

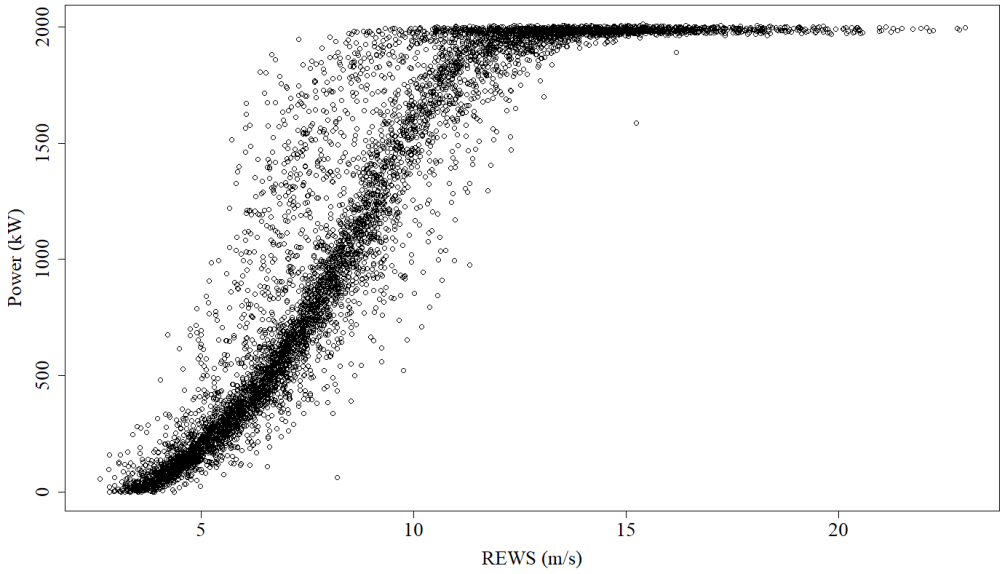

Figure 4: Wind power output vs. rotor equivalent wind speed

REWS represents the speed passing through the entire area swept by a wind turbine's blades into a single value. By considering speed at discrete height intervals, REWS also potentially captures the effects of wind shear and wind veer. These combined effects have led to research into whether it is a better measure of wind speed than HHWS. This is further suggested by observing the differences between Figures 3 and 4. Though

there is still significant spread from the ideal power curve in the REWS plot, the sigmoidal curve nevertheless more closely follows the expected power curve. The deviation from the ideal case is still likely explained by ignoring other factors, but the closer match provides some evidence that REWS may be a better choice for modelling the power curve.

Wind Direction - D

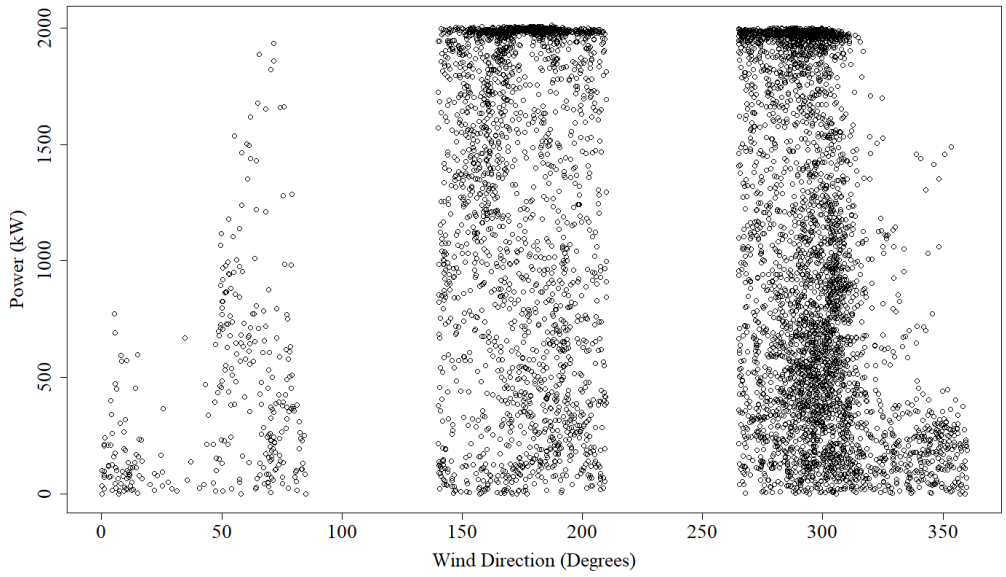

Figure 5: Wind power output vs. wind direction

From prior research (Jeon and Taylor, 2012), it has been determined that wind direction is the most strongly influential variate after wind speed. Intuitively, one can expect that the greatest energy extraction from wind occurs when the wind speed is directed perpendicularly to the face of the area swept by the wind turbine blades. By the same logic, less energy is extracted when wind is directed in parallel to the blades. These observations combine to produce the plot of power against wind direction seen. Clear bands emerge where data are removed in Figure 5. Though wind may blow from any

direction, wind farm supplied datasets tend to omit datapoints associated with the wake effect. The gaps represent the directions that feature another intervening wind turbine, meaning that the resulting power output suffers from the wake effect and is not solely explained by ambient conditions. However, not all wake effect datapoints are necessarily removed. As an example, between the values 145° and 210°, the full scale of power production is represented. By contrast, between the values of 320° and 80°, far fewer data points reach the larger power values, possibly indicating an unfavorable wind direction angle from the wake effect. As a note, wind direction is a circular variable, meaning that a value of 360° wraps around and is equivalent to 0°. As such, 360° and 1° are separated by only a single degree.

Standard Deviation of Wind Direction: σ_D

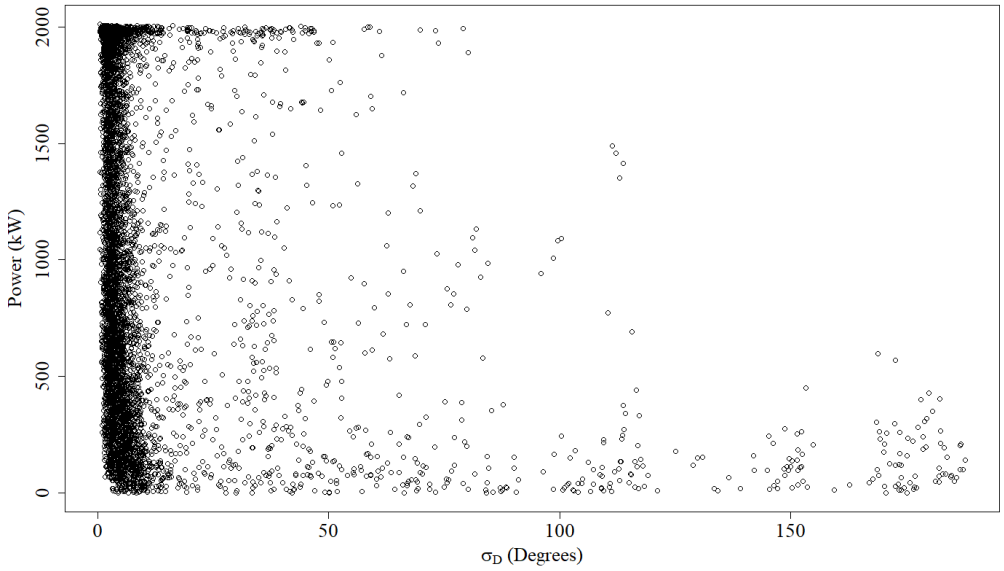

Figure 6: Wind power output vs. standard deviation of wind direction

The standard deviation of wind direction is not directly measured by any onsite equipment, but is instead derived from the measured wind direction. In terms of this dataset, standard deviation at each observation considers the four datapoints before and after as seen in Equation 10 below. Because the dataset is ordered in terms of time increments, this definition of standard deviation measures how much wind direction shifts over a period of time. Low values correspond to stable wind patterns whereas large values correspond to dramatic shifts over short periods of time. From Figure 6, it is apparent that either extreme values of power can be produced for values of σ_D below 50°. However, once σ_D exceeds 50°, there are far fewer datapoints with large values of power production. This indicates that σ_D does not have a significant effect for moderate values, but that large changes in wind direction over a short period of time hamper the turbine's ability to extract energy.

$$\sigma_{D_i} = \sqrt{\frac{1}{9}\sum_{i-4}^{i+4}(x_i - \bar{x})^2} \quad for\ i = 5, 6, 7, \ldots, N-4 \qquad (10)$$

Turbulent Intensity: I

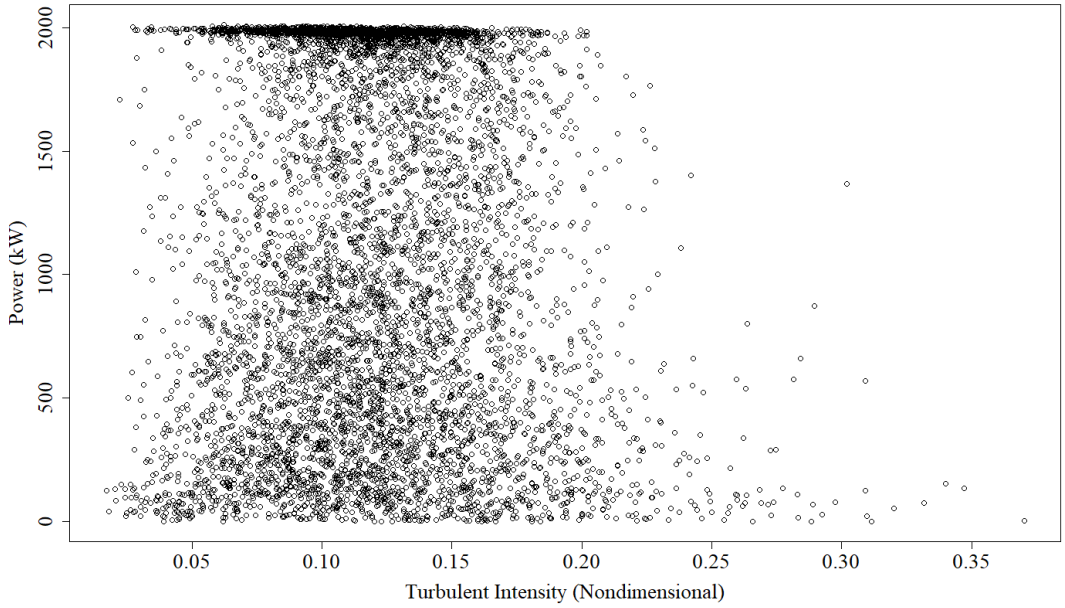

Figure 7: Wind power output vs. turbulent intensity

Turbulent intensity measures how well ordered a wind stream is. Low values correspond to near laminar flow, indicating that the wind is well organized and is pointing in the same direction with the same magnitude across a large area. By contrast, high values correspond with turbulent flow, which involves a variety of wind direction and speed across the area swept by the wind turbine. Figure 7 illustrates the relationship between wind power output and turbulent intensity. For values lower than 0.2, it is apparent that turbulent intensity may not have a strong effect because a wide range of power values are observed. For values greater than 0.2 however, it is notable that fewer observations feature large power values. This may indicate that disorganized, turbulent flows represent a suboptimal scenario for energy extraction.

Clock Time: T

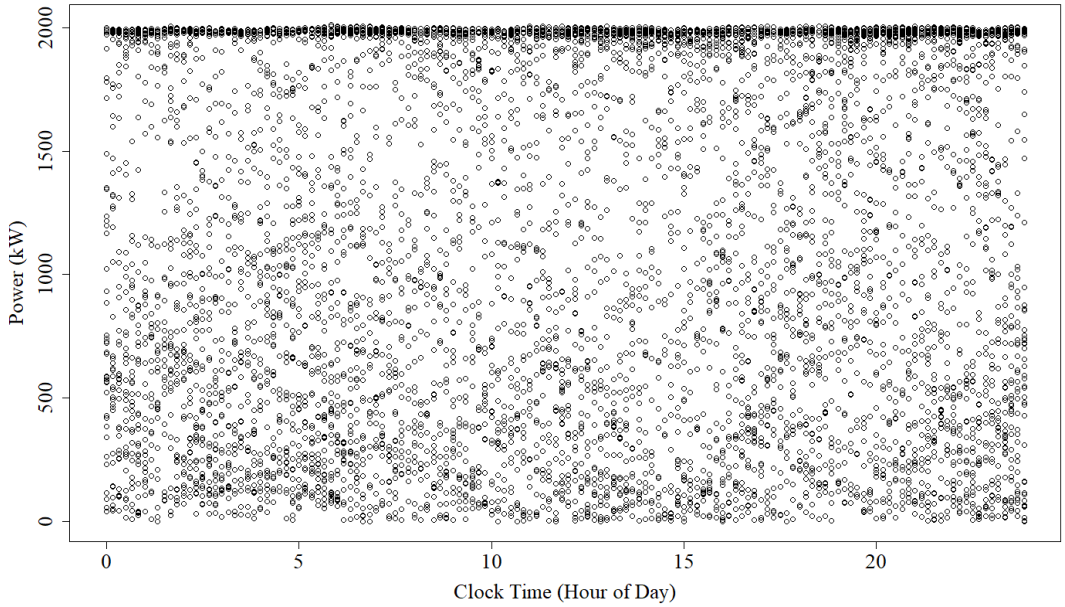

Figure 8: Wind power output vs. clock time

In the original dataset, observations were recorded every 10 minutes for a period of nine months. Each datapoint features a timestamp in addition to the values of the other explanatory variables. Ambient conditions tend to follow seasonal and daily trends, but this research has focused on experimenting with time of day alone. As an example, both wind speed and temperature tend to follow day/night cycles that may be captured in a model including time as an input. For use in the AMK model, the clock time was converted from HH:MM format to real, values in hours. However, no immediate trend emerges from Figure 8 showing the relationship between power and clock time. Though the subsequent experimentation will clarify the use of including time, it may be necessary to separate out seasonal trends in order for daily effects to be apparent.

Wind Shear: S

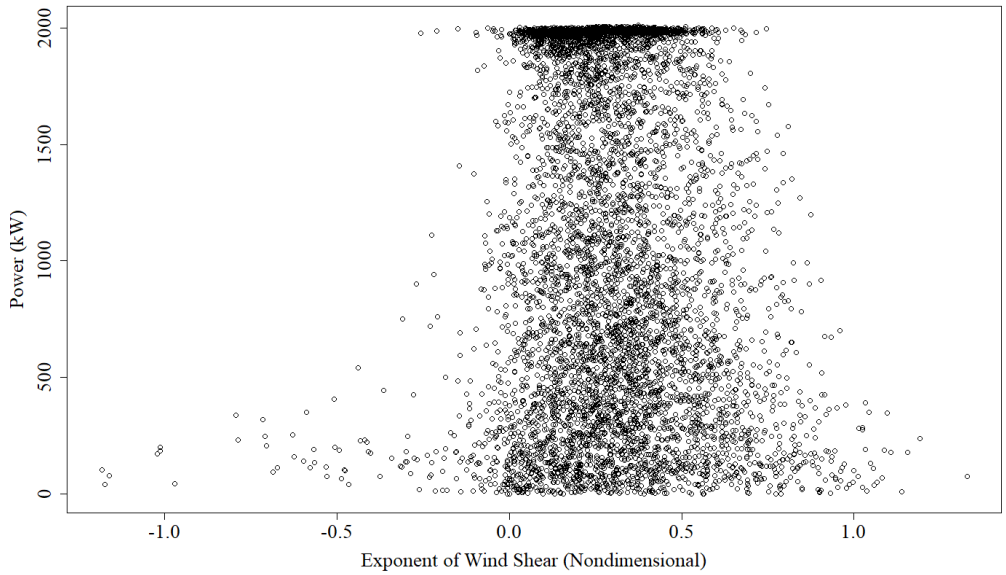

Figure 9: Wind power output vs. exponent of wind shear

Wind shear is another derived statistic that measures how much wind speed varies at different elevations. It essentially captures the speed gradient along the height of the area swept by the wind turbine blades. Though it can be expressed with the same units as speed, the unitless wind shear exponent is instead calculated here according to Equations 11 and 12 below. This expression is defined due to the no-slip condition, which states that fluids features zero velocity at solid surfaces. The speed profile then follows an exponential relationship to the distance from said surface. In this work, wind shear is calculated separately for the areas above and below the turbine's hub. Equation 11 corresponds to the upper wind shear, whereas equation 12 refers to the lower portion. Positive values of the exponent indicate that speed increases as height increases, whereas a negative value indicates that speed decreases as height increases. From Figure 9, it is clear that fewer

observations have large power values as the exponent deviates from a central tendency of 0.25. Figure 9 shows the relationship for the upper portion of wind shear, but the relationship for the lower portion follows a nearly identical plot.

$$S_a = \frac{\ln \frac{V_T}{V_H}}{\ln \frac{Z_T}{Z_H}} \qquad (11)$$

$$S_b = \frac{\ln \frac{V_H}{V_B}}{\ln \frac{Z_H}{Z_B}} \qquad (12)$$

Wind Veer: E

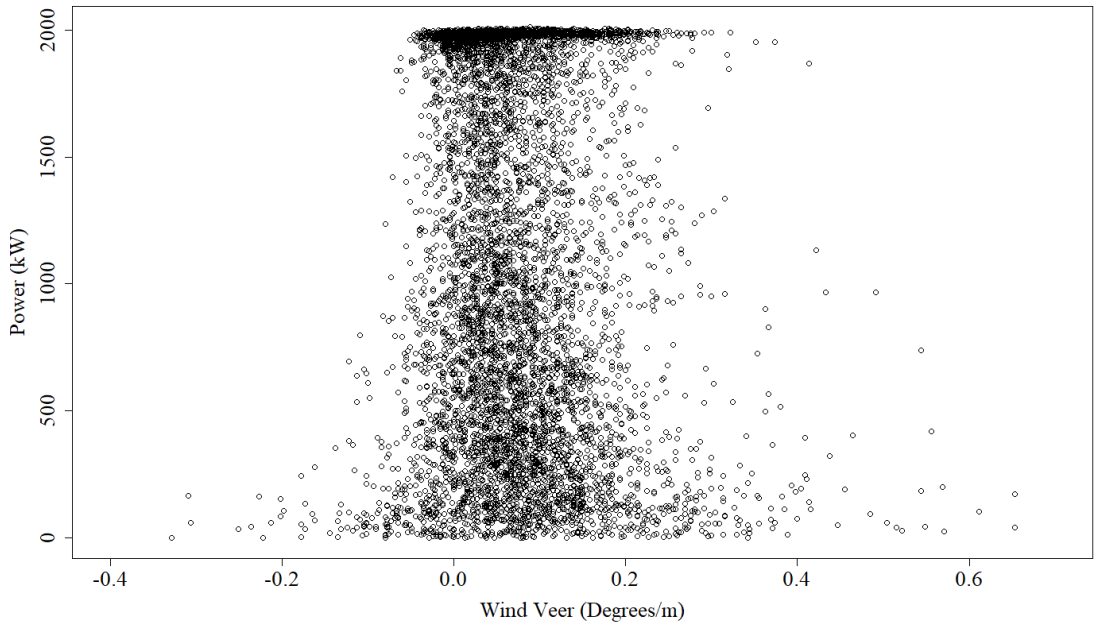

Figure 10: Wind power output vs. wind veer

Wind veer is similarly defined to wind shear, but instead measures how wind direction varies along the height of the wind turbine. Rather than be expressed as an

exponent, wind veer is directly represented as a ratio of the difference of wind directions and their respective heights. Equations 13 and 14 express wind veer as follows for cases above and below the hub respectively. As in wind shear, positive values indicate that wind direction increases in value as height increases, whereas negative values mean that wind direction decreases in value as height increases. The relationship between wind veer and power appears similar to that of wind shear. From Figure 10, lower power values result as wind veer deviates from a central tendency of 0.1. When wind veer is approximately at 0.1 however, the full range of power values are observed.

$$E_a = \frac{\Delta D}{\Delta Z} = \frac{D_T - D_H}{Z_T - Z_H} \qquad (13)$$

$$E_b = \frac{\Delta D}{\Delta Z} = \frac{D_H - D_B}{Z_H - Z_B} \qquad (14)$$

3.2 Objective and Performance Metrics

The objective of the present work is to explore the variable space and demonstrate empirically which combination of factors leads to the greatest error reduction. As noted previously, consensus has not yet emerged as to which set of factors should necessarily be included in modelling the power curve. By trialing different combinations of explanatory variables and evaluating error when implemented into the AMK model, evidence is gathered on which factors should be included in future models. These trials will be ranked by prediction error to select the best model.

Permutations of the factors are attempted and evaluated in terms of root mean square error (RMSE) and mean absolute error (MAE) compared against a known response. The final deliverable consists of a table of RMSE and MAE values for each combination attempted. The expressions for RMSE and MAE follow in Equations 15 and 16. N_{TS} refers to the testing set used over which $m(x_i)$ is estimated, and y_i is the known testing response.

$$RMSE = \sqrt{\frac{1}{N_{TS}} \sum_{i=1}^{N_{TS}} (\hat{m}(x_i) - y_i)^2} \qquad (15)$$

$$MAE = \frac{1}{N_{TS}} \sum_{i=1}^{N_{TS}} |\hat{m}(x_i) - y_i| \qquad (16)$$

Both RMSE and MAE are included to measure error in order to capture different facets of error. The error in RMSE grows quadratically as the prediction deviates from the actual values. As such, prediction outliers are heavily penalized, whereas predictions close to the actual values enjoy a very mild penalty. By contrast, the error expressed by MAE grows linearly as the prediction deviates from the actual value. Outliers are less penalized than in RMSE, but predictions close to the actual value suffer a heavier penalty comparatively. Differences in reduction of error in terms of RMSE and MAE will indicate how well models reduce the number of outlier predictions in the case of RMSE versus how well they reduce overall error in the case of MAE.

3.3 Outline of Numerical Analysis Procedure

3.3.1 Overall Description

The overarching goal of this research project is to select the best model of AMK to use in this particular dataset, and to determine which set of variables achieves this goal. This work of research ranks competing sets of models through error values outputted in a series of trials. Each trial consists of solving the power curve problem using a specific set of explanatory variables input into AMK as the learning method. 10-fold cross-validation is used, meaning that the original dataset is randomly split into ten training and testing sets, where ten sets of power predictions are made and compared against the actual response in order to generate ten values of RMSE and MAE each. These values are averaged to output the final RMSE and MAE values associated with that trial's set of explanatory variables. The goal is to select the three models that feature the lowest RMSE and MAE values to gain insight into which variables are of interest to the power curve problem.

As to which variables to experiment with, prior work with AMK has already shown that wind speed and direction should always be included (Lee et al., 2015). These factors not only strongly impact power production, but also feature interaction effects with other variables. Because of this, wind speed and wind direction are always used with other variables interchanged when running trials. The remaining sets of variables can be split into two categories. The first involve variables that measure wind speed and direction gradients with respect to height. As the size of turbine blades grows, there is concern that single-point measurements will not suffice for use in modelling the power curve. As such, the previously mentioned REWS has been introduced compared to the historical HHWS. Additionally, wind shear and wind veer can also be explicitly inputted into learning

models. Evidence as to which to use is provided by the results of this research. The second category involves miscellaneous variables typically captured by wind turbine datasets. Turbulent intensity, the standard deviation of wind direction, and the time of day are explored. The best models resulting from either category are ultimately combined to find the overall best model.

3.3.2 Code Description

As of the date of writing, no standard R package of AMK exists for download. Instead, an implementation of the AMK method was first reproduced in R Studio prior to commencing the study. The code allows for interchange of input datasets as need be. Bandwidths for all explanatory variables are calculated via the "dpill" package. As mentioned before, the bands apparent in wind direction and the standard deviation of wind direction require that two distinct bandwidth values be calculated for each of these cases. Before finally running the learning method, the original dataset is subjected to a ten-part split to facilitate 10-fold cross-validation. With these preliminary steps complete, the Nadaraya-Watson estimates are computed with the input of the training set, testing set, and bandwidth values. The code runs through all 37 subsets of variables trialed and iterated ten times. Finally, RMSE and MAE values are computed from the average error values across the ten sets for each of the 37 trials.

3.3.3 Sequence of Trials

Gradient Trials

The first set of trials involves comparing REWS to HHWS, as well as determining whether the addition of wind shear and veer lead to more reduction of error. Wind speed, wind direction, and air density are always included in these trials. Following the variable

notation established in the dataset section above, these three variables form the *(V,D,A)* base to compute the mean conditional density estimator.

$$\hat{m} = \hat{m}(V_H, D, A) \qquad (17)$$

AMK can accept a full model of all explanatory variables, but only considers three variables in each additive term as shown in Equation 17. To determine the effect of the gradient effect, it is necessary to also include one of the variables associated with the gradient effect into the model. This additional variable will take the place of air density in a second additive term as shown in Equation 18.

$$\hat{m} = \frac{1}{2}[\hat{m}(V_H, D, A) + \hat{m}(V_H, D, S_a)] \qquad (18)$$

There will be as many additive terms as there are explanatory variables in a given model. The additional gradient effect variables considered are shown in the set below. They can each be added individually or in combination to the AMK model.

$$\{V_H, V_R, S_a, S_b, E_a, E_b\}$$

The last four terms refer to wind shear and wind veer, both above and below the hub height. These factors take the place of the third variable in an AMK additive term. In addition, the speed term in Equation 18 is also subject to experimentation, taking as its value the REWS or HHWS version. Including both in a single model is also considered. A total of eighteen trials result from these combinations. These trials serve to isolate the effect of including the gradient of wind speed and direction in the AMK model. The results of the trials in this section will answer whether REWS or HHWS better expresses wind speed. It will also become apparent whether they can both be considered at once to avoid the issue of selecting either over the other. Finally, the trials involving wind shear and

wind veer determine whether including REWS is redundant if the gradient effect is better captured through these variables.

Miscellaneous Trials

The second set of trials involves the remaining explanatory variables included in the PCWG dataset, as well as the standard deviation of wind direction. This set includes the following variables. Each factor is used as the third variable in each AMK additive term.

$$\{T, \sigma_D, I\}$$

As in the gradient trials, it is first necessary to isolate these variables to measure their baseline effect. With that established, combinations of these variables are attempted. The gradient variables are excluded for the moment, and are combined with the miscellaneous factors in the next set of trials. As such, only V_H is considered in this section to serve as a baseline for further improvement. (V, D) and (V, D, A) were used as the initial bases, with combinations of the three explanatory variables in question then attempted. A total of ten trials result from this section. From this section, it becomes apparent which of these variables leads to the greatest reduction of error.

Combined Trials

The third set of trials involves combining the best models from the preceding two sections to determine if additional reduction in error is attainable. Specifically, the three best models from the miscellaneous trials are selected. One of three modifications are made to each of these models to add the gradient effects. The first modification replicates the best models using V_R instead of of V_H. S_a and E_a are then added to each of the best

models from the miscellaneous trials. Using either V_H and V_R in these cases represent the remaining two modifications. A total of nine trials result from these combinations.

4. RESULTS

A few key results follow from the experimentation with the explanatory variables. The RMSE and MAE values are tabulated in the first section below, and is presented with discussion. Finally, figures showing the predicted power curve for the best models are compared to the actual values.

4.1 Discussion

4.1.1 Gradient Trials

The gradient trials show that REWS performs better than HHWS in all cases. However, using both REWS and HHWS in the same model significantly increases the error of the model. This is consistently true in Tables 1, 2, and 3. Including wind sheer and wind veer leads to some marginal improvement, however, their inclusion in future models remains debatable. The addition of these variables into AMK increases model complexity without capturing information not already used to determine REWS. With such a modest decrease in error, it may be better to exclude wind shear and wind veer to avoid the risk of overfitting. Doing so may provide better results on generating estimates on datasets not used in training the model. The tables in this section are split into sections according to what set of explanatory variables are used. For each set, V_H, V_R, or both are used for the speed term.

Gradient Trials - Set 1							
	(V, D, A)			**(V, D, A, S_a)**			
	V_H	V_R	V_HV_R	V_H	V_R	V_HV_R	
RMSE	102.37	99.70	140.31	99.22	98.20	156.94	*RMSE*
MAE	64.12	62.21	87.90	63.32	62.14	97.36	*MAE*

Table 1: Gradient Trials - Set 1

Gradient Trials - Set 2							
	(V, D, A, S_b)			**(V, D, A, E_a)**			
	V_H	V_R	V_HV_R	V_H	V_R	V_HV_R	
RMSE	102.86	100.19	159.61	102.60	100.29	161.39	*RMSE*
MAE	64.81	62.91	97.69	65.13	63.25	99.09	*MAE*

Table 2: Gradient Trials - Set 2

Gradient Trials - Set 3							
	(V, D, A, E_b)			**(V, D, A, S_a, E_a)**			
	V_H	V_R	V_HV_R	V_H	V_R	V_HV_R	
RMSE	105.78	102.70	161.33	99.15	97.96	169.64	*RMSE*
MAE	65.84	63.84	98.81	63.40	62.08	103.88	*MAE*

Table 3: Gradient Trials - Set 3

4.1.2 Miscellaneous Trials

The miscellaneous trials reveal that the inclusion of the time and turbulent intensity lead to a reduction of error, while the standard deviation of wind direction leads to a slight increase compared to the (V_H, D, A) base model. Table 4 shows the individual effect of including turbulent intensity, the standard deviation of wind direction, and time. Table 5 gives the error values for models combining those factors.

Even after separating the σ_D domain into two intervals and calculating a corresponding bandwidth for both, the resulting model does not positively impact the prediction error. Experimentation was done to select the best partition point for the σ_D domain, so it is doubtful whether improvements can be made by including σ_D in modelling the power curve in this case. The turbulent intensity results by contrast demonstrate that the turbulence of incoming wind does impact a turbine's ability to extract energy. It is also apparent that the time of day has some influence, likely because wind patterns follow a daily cyclical trend. Because both of these variables depend on the nature of incoming

wind, it is likely that the AMK model is capturing an interaction effect between them and the wind speed and direction. Due to the noted reduction of error, both turbulent intensity and time are included in the subsequent combined models to measure their interaction with the gradient effects.

Miscellaneous Trials - Individual Variable Effects						
	Intensity: I		**Std.Dev. of D: σ_D**		**Time: T**	
	(V_H, D, I)	*(V_H, D, A, I)*	*(V_H, D, σ_D)*	*(V_H, D, A, σ_D)*	*(V_H, D, T)*	*(V_H, D, A, T)*
RMSE	105.21	98.65	125.58	107.42	102.10	96.72
MAE	66.16	62.18	73.19	66.55	64.12	61.24

Table 4: Miscellaneous Trials - Individual Variable Effects

Miscellaneous Trials - Combined Variable Effects					
	VDA Base: (V_H,D,A,…)				
	(V_H, D, A, I, σ_D)	*(V_H, D, A, I, T)*	*(V_H, D, A, T, σ_D)*	*(V_H, D, A, I, T, σ_D)*	
RMSE	103.65	95.16	154.23	133.56	*RMSE*
MAE	65.15	60.35	99.33	86.65	*MAE*

Table 5: Miscellaneous Trials - Combined Variable Effects

4.1.3 Combined Trials

Combining models from the gradient and miscellaneous trials leads to further reduction in error. Three base models are considered here by taking the best models from the miscellaneous trials and adding the gradient effect. Tables 6, 7, and 8 correspond to the *(V, D, A, I)*, *(V, D, A, T),* and *(V, D,A ,I, T)* base models respectively. The best result comes from the *(V_R, D, A, I, T)* model.

The combination of REWS with time and turbulent intensity leads to an approximately 8.6% and 8% drop in RMSE and MAE respectively compared to that of the *(V, D, A)* model. Also including wind shear and wind veer led to a similar RMSE but a 7.1% drop in MAE. As in the gradient trials however, the error reduction by including

wind shear and wind veer must be weighed against the potential to overfit models. This is especially true considering that REWS essentially captures the same information as wind shear and wind veer. Further testing may be necessary to conclude whether wind shear and veer lead to improved predictions on a new year of data at this wind farm.

Combined Trials - Set 1					
	(V,D,A,I)		(V_H,D,A,I,S_a,E_a)		
	V_H	V_R	V_H	V_R	
RMSE	98.65	96.50	97.37	96.22	*RMSE*
MAE	62.18	60.64	62.22	60.97	*MAE*

Table 6: Combined Trials - Set 1

Combined Trials - Set 2					
	(V,D,A,T)		(V_H,D,A,T,S_a,E_a)		
	V_H	V_R	V_H	V_R	
RMSE	96.72	94.98	95.54	94.50	*RMSE*
MAE	61.24	59.79	61.29	60.11	*MAE*

Table 7: Combined Trials - Set 2

Combined Trials - Set 3					
	(V,D,A,I,T)		(V_H,D,A,I,T,S_a,E_a)		
	V_H	V_R	V_H	V_R	
RMSE	95.16	93.52	94.85	93.80	*RMSE*
MAE	60.35	58.98	60.77	59.59	*MAE*

Table 8: Combined Trials - Set 3

4.1.4 Ranked Trials

Finally, the best models from the three preceding sections are presented here to clearly show the models with the best performance in terms of error reduction. Table 9 shows the best four models, and Table 10 shows the next four best models. Error reduction in terms of percentage change compared to the base *(V_H, D, A)* model is also given to measure error reduction resulting from using additional variables. It is important to note

that Table 9 features models using V_R exclusively because they outperformed their corresponding V_H counterparts. This result suggests that REWS should be used in place of HHWS for the speed term in power curve modelling.

Ranked Models - REWS					
	(V_H,D,A)	(V_R,D,A,I,T)	(V_R,D,A,I,T,S_a,V_a)	(V_R,D,A,T)	(V_R,D,A,T,S_a,E_a)
RMSE	102.37	93.52	93.80	94.98	94.50
% Change	Baseline	-8.6	-8.4	-7.2	-7.7
MAE	64.12	58.98	59.59	59.79	60.11
% Change	Baseline	-8.0	-7.1	-6.8	-6.3

Table 9: Ranked Models – REWS

Ranked Models - HHWS					
	(V_H,D,A)	(V_H,D,A,I,T)	(V_H,D,A,I,T,S_a,E_a)	(V_H,D,A,T,S_a,E_a)	(V_H,D,A,T)
RMSE	102.37	95.16	94.85	95.54	96.72
% Change	Baseline	-7.0	-7.4	-6.7	-5.5
MAE	64.12	60.35	60.77	61.29	61.24
% Change	Baseline	-5.9	-5.2	-4.4	-4.5

Table 10: Ranked Models – HHWS

4.2 Predicted Power Curve Examples

The figures below demonstrate the power curve estimates generated by the three best models from Table 9 ranked by increasing RMSE. These predicted curves are compared against the actual power curve from the testing set. These models were found by comparing error values after performing 10-fold cross-validation, but the curves below instead use a 90/10 split for training and testing respectively. A model is trained on 90% of the original PCWG dataset, and then predicts power values based on the explanatory variables of the remaining 10% of the dataset.

Figures 11, 12, and 13 show the *(V_R, D, A, I, T)*, *(V_R, D, A, I, T, S_a, E_a)*, and *(V_R, D, A, T)* models respectively each compared against the actual power curve with REWS chosen as the speed measure. The black circles represent the testing set power values, whereas the blue, green, and red circles correspond to each model's power estimates.

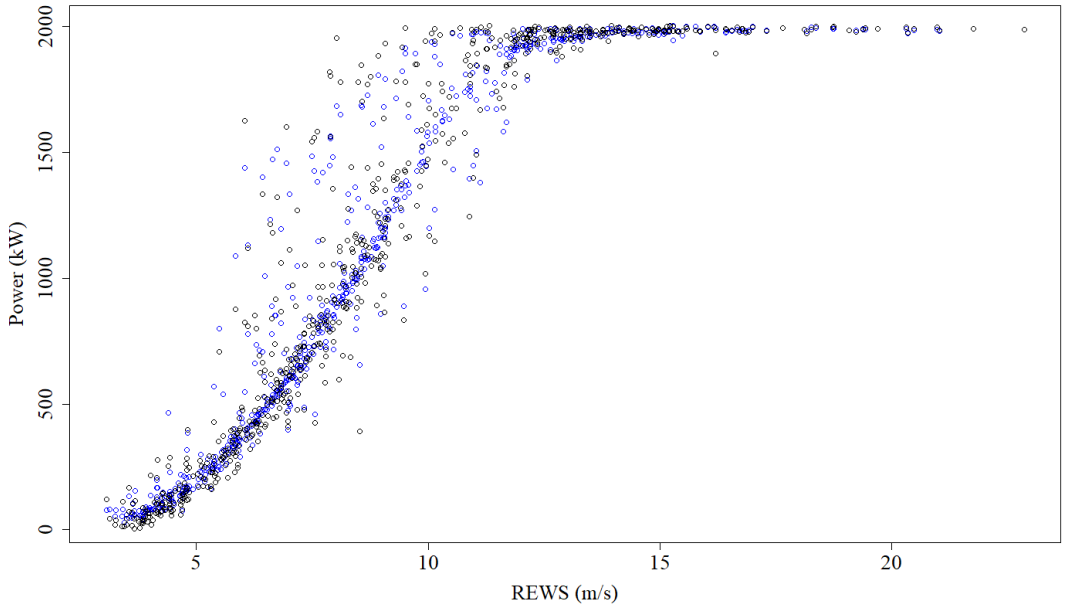

Figure 11: Estimated vs. actual power curve - (V_R, D, A, I, T) model

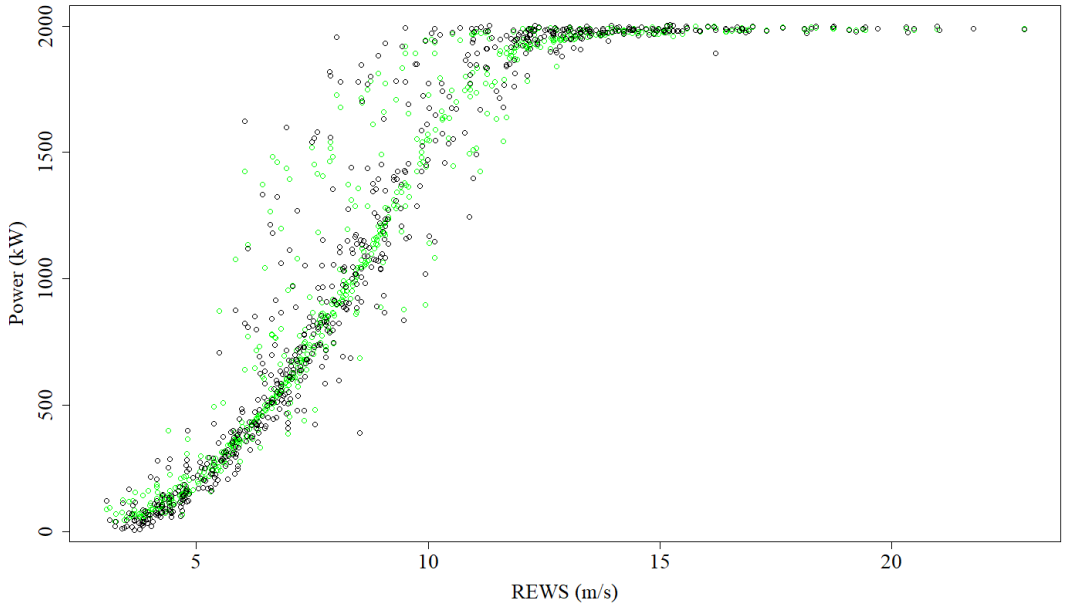

Figure 12: Estimated vs. actual power curve - (V_R, D, A, I, T, S_a, E_a) model

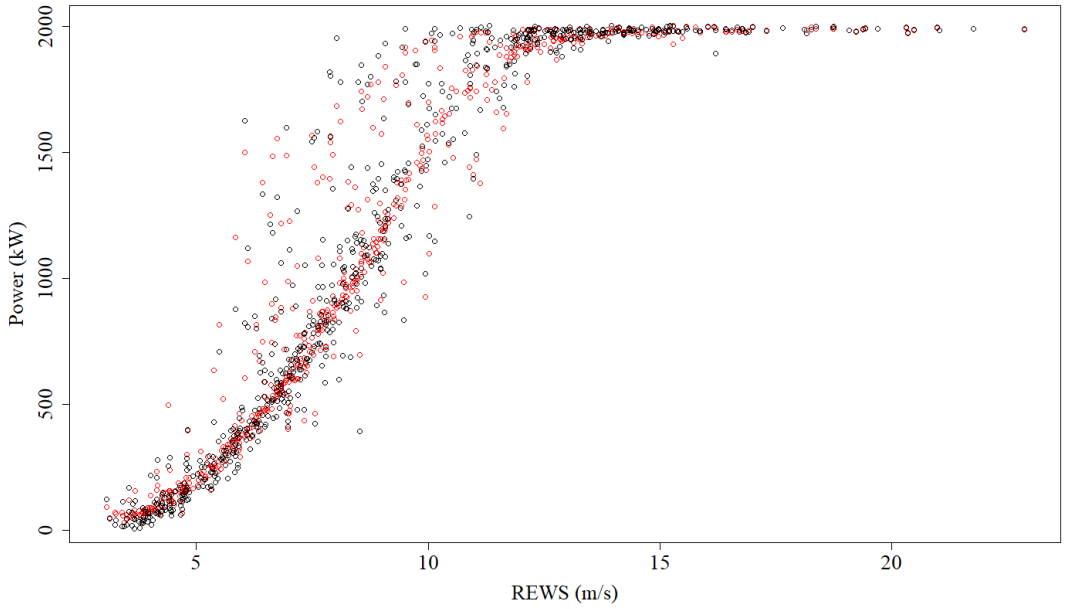

Figure 13: Estimated vs. actual power curve - (V_R, D, A, T) model

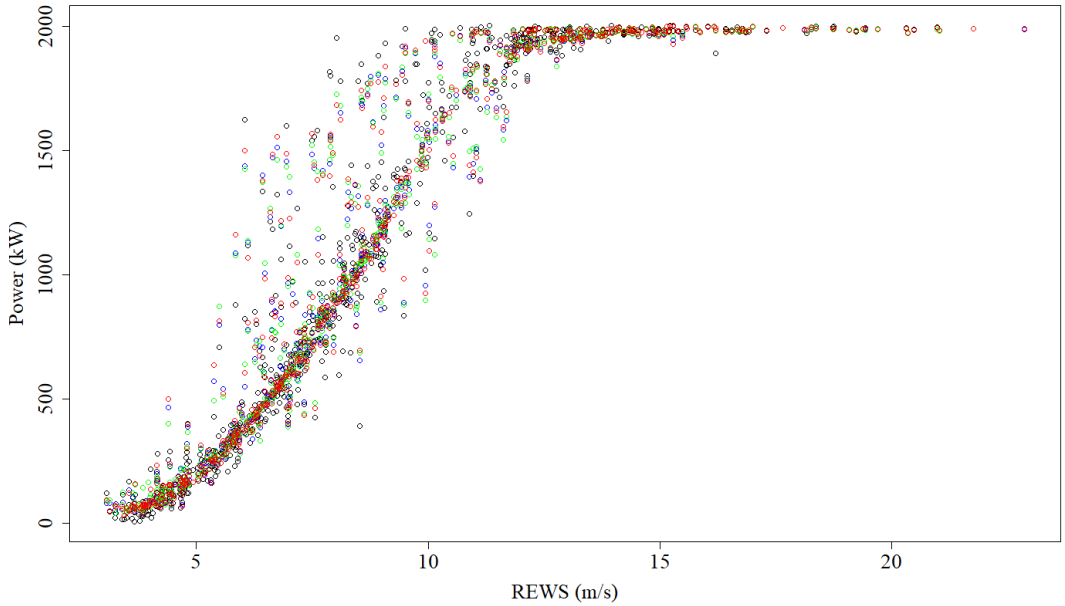

Figure 14: Estimated vs. actual power curve - best models

Figure 14 shows all three models' power estimates in one plot to highlight differences among each. The power curve deviates most strongly from a nominal sigmoidal curve for values of REWS between 6 and 11 m/s. The power curve estimates strongly follow the actual power curve outside of this speed range, but suffer from deviation from the actual values and a nominal sigmoidal curve inside that speed range. Though the estimates deviate from actual values, the estimated curves demonstrate they follow the pattern of the actual power output values.

5. CONCLUSION

This work of research attempts to provide a foundation on how to approach new power curve datasets while setting a site-specific methodology to explore the variable space to determine the best possible model. Two important results can be gathered from this problem. The first is the general methodology to approach this type of problem. Division of the subsets of variables used in the experimental trials was done with understanding of the nature of the explanatory variables and their relationship with the power response. Secondly, strong evidence is provided that the rotor equivalent wind speed outperforms hub height wind speed in the power curve problem. The best rotor equivalent wind speed model is able to outperform the base hub height wind speed model by nearly 9% in terms of root mean square error reduction. In addition, all models that included rotor equivalent wind speed performed better than their counterpart models that instead used hub height wind speed. As such, evidence is added to the hypothesis that the wind industry needs to capture gradient effects. This is especially true given the size of contemporary wind turbines compared to older models. The inclusion of time and turbulent intensity also lead to consistent improvement in nearly all models, indicating that their influence on power generation is substantial enough to justify their use in future modelling efforts.

REFERENCES

Belghazi, O., & Cherkaoui, M. (2012). Pitch angle control for variable speed wind turbines using genetic algorithm controller. *Journal of Theoretical and Applied Information Technology*, 39, 5-10.

Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., & Draxl, C. (2011). The state-of-the-art in short-term prediction of wind power: A literature overview (2nd ed). ANEMOS. Retrieved from https://pdfs.semanticscholar.org/45e8/3e5fac2bfa0bf999f3ec596302afae43cda4.pdf

Hyndman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, 5, 315-336.

IEC (2005), IEC 61400-12-1 Ed 1, *Wind Turbines-Part 12-1: Power Performance Measurements of Electricity Producing Wind Turbines*, Geneva, Switzerland: International Electrotechnical Commission.

Jeon, J., & Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association*, 107, 66-79.

Lee, G., Ding, Y., Genton, M. G., & Xie, L. (2015). Power curve estimation with multivariate environmental factors for inland and offshore wind farms. *Journal of the American Statistical Association*, 110, 56-67.

Power Curve Working Group, www.pcwg.org, [last accessed on January 17, 2018]

Rosenblatt, M. (1969). Conditional probability density and regression estimators. *Multivariate Analysis II*, 25-31.

Ruppert, D., Sheather, S. J., & Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90, 1257-1270.

Scheurich, F., Enevoldsen, P. B., Paulsen, H. N., Dickow, K. K., Fiedel, M., Loeven, A., & Antoniou, I. (2016, September). Improving the accuracy of wind turbine power curve validation by the rotor equivalent wind speed concept. *Journal of Physics: Conference Series*, 753.

Uluyol, O., Parthasarathy, G., Foslien, W., & Kim, K. (2011). Power curve analytic for wind turbine performance monitoring and prognostics. *Annual conference of the prognostics and health management society*, 2, 1-8.